Alberta Thy 15-02

Towards a standard jet definition

D. Yu. Grigoriev[1,3], E. Jankowski[2], and F. V. Tkachov[3]
[1] *Mathematical Physics, Natl. Univ. of Ireland Maynooth, Maynooth, Co. Kildare, Ireland*
[2] *Department of Physics, University of Alberta, Edmonton, AB, T6G 2J1, Canada*
[3] *Institute for Nuclear Research of RAS, Moscow 117312, Russia*

(Dated: June 4, 2005)

In a simulated measurement of the W-boson mass, evaluation of Fisher's information shows the optimal jet definition [1] to yield the same precision as the k_{T} algorithm while being much faster at large multiplicities.

1. Association of hadronic jets observed in high energy physics experiments with quarks and gluons in the underlying collisions of quanta [2] provides an experimental handle on fundamental interactions via the so-called jet finding algorithms that find a configuration of jets $\mathbf{Q}$, represented by the N_{jets} 4-momenta p_j, for a given event $\mathbf{P}$, represented by the N_{part} light-like 4-momenta p_a:

$$\mathbf{P} = \{p_a\} \xrightarrow{\text{jet algorithm}} \mathbf{Q} = \{p_j\}\,. \tag{1}$$

Unless jets are energetic and well separated, jet definition involves ambiguities that were seen to be a major, even dominant source of errors in the planned experiments [3].

A well-known requirement on possible jet finding algorithms is the infrared safety [4] or insensitivity of $\mathbf{Q}$ to collinear fragmentations of particles in $\mathbf{P}$. It is clarified by a theorem of ref. [5] expressing fragmentation-invariant observables in terms of the energy-momentum tensor defined by space-time symmetries uniquely, so that such observables can be equivalently represented in terms of either hadron, or quark and gluon fields. However, the requirement leaves much freedom for the mapping (1), and many jet algorithms emerged over time.

2. Ref. [4] introduced so-called cone algorithms that define a jet as all particles in a cone of a fixed radius [6]. Cone axes are usually found iteratively to be directed along jets' 3-momenta, and cone overlaps are treated with ad hoc prescriptions. The fixed shape of cones enhances the stability of cone algorithms and facilitates studies of detector corrections, but decreases the jet resolution power.

Ref. [7] introduced a definition based on the shape observable thrust [8], as theoretical studies are easier with such observables. Here one minimizes the sum

$$\textstyle\sum_j \left(1 - T_j\right), \tag{2}$$

where T_j is the thrust for the j-th jet. (Similar measures were considered e.g. in [32] and a special case of jet search based on an optimization of a shape observable was also employed e.g. in [33].) However, the required minimization was deemed unfeasible [9].

Successive recombination algorithms emerged with a motivation to invert hadronization [10]. Here one starts with a list of particles, computes a "distance" d_{ab} for each pair of particles, and replaces the pair with the smallest d_{ab} by a single pseudo-particle with $p_{ab} = p_a + p_b$. One repeats this until e.g. all d_{ab} exceed a given threshold y_{cut} or only a given number of (pseudo-) particles remain in the list. Possible d_{ab} are given by

$$d_{ab}^2 = E_a E_b \left(E_a + E_b\right)^{-n} \left(1 - \cos\theta_{ab}\right), \tag{3}$$

where E_a and E_b are the particles' energy fractions, θ_{ab} is the angle between them, and y_{cut} is the so-called jet resolution parameter. $n = 0$ and $n = 2$ correspond to the JADE [11] and Geneva [12] criteria. It was also argued that the dynamics of the $2 \to 1$ amplitude in QCD is matched best by the so-called k_{T} measure [13]:

$$d_{ab}^2 = \min\left(E_a^2, E_b^2\right)\left(1 - \cos\theta_{ab}\right). \tag{4}$$

Such algorithms find jets of irregular shapes. Ref. [14] replaced $2 \to 1$ recombinations with a global $n \to m$ one (but still based on pairwise distances d_{ab}), yielding more regular jets but this is more expensive computationally.

The multitude of available jet algorithms — often differing in obscure details — caused their comparative studies (e.g. refs. [6], [12], [10], [15]). The subject's importance has been growing along with the drive towards higher precision in the jet physics [3], [15].

3. Ref. [16] reinterpreted the physically significant ambiguities of jet algorithms due to algorithmic variations as instabilities which a correct measurement procedure must be free of. The resulting theory [1], [17] provided a context to derive an optimal jet definition from explicit physical motivations. The principal points of the theory are as follows:

(i) Calorimetric measurements with hadronic final states $\mathbf{P}$ must rely on observables $f(\mathbf{P})$ that possess a special "calorimetric", or C-continuity which is a nonperturbative generalization of the familiar IR safety (see [17] for details) and which guarantees a stability of $f(\mathbf{P})$ against distortions of $\mathbf{P}$ such as caused by detectors. Ref. [17] pointed out C-continuous analogues for a variety of observables usually studied via intermediacy of jet algorithms. The fundamental role of such observables is highlighted by two facts: (1) An observable inspired by [17] played an important role in the selection of top quark events in the fully hadronic channel at D0 [19], [20]. (2) The Jet Energy Flow project [21] provides numerical evidence that C-continuous observables may indeed help to go beyond the intrinsic limitations of conventional procedure based on jet algorithms in the quest for the 1% precision level in the physics of jets.

(ii) The proposition that the observed event $\mathbf{P}$ inherits information (as measured by calorimetric detectors) from the underlying quark-and-gluon event $\mathbf{q}$ is expressed as

$$f(\mathbf{q}) \approx f(\mathbf{P}) \quad \text{for any } C\text{–continuous } f. \tag{5}$$

(iii) For each parameter M on which the probability distribution $\pi_M(\mathbf{P})$ of the observed events $\mathbf{P}$ may depend, there exists an optimal observable $f_{\text{opt}}(\mathbf{P}) = \partial_M \ln \pi_M(\mathbf{P})$ for the best possible measurement of M [18]. This is a reinterpretation of the Rao-Cramer inequality and the maximal likelihood method of mathematical statistics in terms of the method of moments. In the context of multi-hadron final states as "seen" by calorimetric detectors, such an observable is automatically C-continuous.

(iv) If the dynamics of hadronization is such that eq. (5) holds, then good approximations for f_{opt} could exist among functions that depend only on $\mathbf{Q}$ which is a parameterization of $\mathbf{P}$ in terms of a few pseudo-particles (jets), found from a condition modeled after eq. (5):

$$f(\mathbf{Q}) \approx f(\mathbf{P}) \quad \text{for any } C\text{–continuous } f. \tag{6}$$

This simply translates the meaning of jet finding as an inversion of hadronization into the language of C-continuous observables.

(v) C-continuous observables can be approximated by sums of products of simplest such observables that are linear in particles' energies:

$$f(\mathbf{P}) = \sum_a E_a f(\hat{\mathbf{p}}_a). \tag{7}$$

(The relevant theorems can be found in refs. [1] and [17].)

(vi) So it is sufficient to explore the criterion (6) with only f's of the form (7). Then one can perform a Taylor expansion in angular variables and obtain a factorized bound of the form

$$|f(\mathbf{P}) - f(\mathbf{Q})| \leq C_{f,R} \times \Omega_R[\mathbf{P},\mathbf{Q}], \tag{8}$$

where all the dependence on f is localized within $C_{f,R}$ (so the bound remains valid for any C-continuous f) and

$$\Omega_R[\mathbf{P},\mathbf{Q}] = R^{-2}\Upsilon[\mathbf{P},\mathbf{Q}] + \mathrm{E}_{\text{soft}}[\mathbf{P},\mathbf{Q}], \tag{9}$$

where $\Upsilon[\mathbf{P},\mathbf{Q}] = 2\sum_j p_j \tilde{q}_j$, $\mathrm{E}_{\text{soft}}[\mathbf{P},\mathbf{Q}] = \sum_a \bar{z}_a E_a$, and $R > 0$ is a free parameter (see ref. [1] for a discussion). p_j are jets' physical 4-momenta expressed as $p_j = \sum_a z_{aj} p_a$, where the so-called recombination matrix z_{aj} is such that $0 \leq z_{aj} \leq 1$ and $\bar{z}_a = 1 - \sum_j z_{aj} \geq 0$ for any a, i.e. a part of the particle's energy is allowed to not participate in any jet. $\tilde{q}_j$ are light-like 4-vectors related to p_j and given by $\tilde{q}_j = (1, \mathbf{p}_j/|\mathbf{p}_j|)$ for lepton collisions ($\tilde{q}_j$ can be definied differently for hadron collisions; see ref. [1] for details). The recombination matrix z_{aj} occurs naturally in the construction of the bound (8) and is the fundamental unknown in this scheme. Υ in (9) differs from (2) in that the jet's physical momentum is used in place of the thrust axis. E_{soft} is the event's energy fraction that does not take part in jet formation. (vii) Since the collection of values of all f on a given event $\mathbf{P}$ is naturally interpreted as the event's physical information content, the bound (8) means that the distortion of such content in the transition from $\mathbf{P}$ to $\mathbf{Q}$ can be controlled by a single function; so the loss of physical information in the transition is minimized if $\mathbf{Q}$ corresponds to the global minimum of Ω_R. The Optimal Jet Definition amounts to finding z_{aj} which minimizes Ω_R, depending on specific application, either with a given number of jets or with a minimum number of jets while satisfying the restriction $\Omega_R[\mathbf{P},\mathbf{Q}] < \omega_{\text{cut}}$ with some parameter $\omega_{\text{cut}} > 0$ which is similar to the jet resolution y_{cut} of recombination algorithms.

4. OJD combines attractive features of the different algorithms reviewed above and is free of their defects (see ref. [1] for more details): (i) OJD is based on a shape observable. (ii) It finds jets of rather regular shapes with angular radii bounded by R. (iii) it resolves jet overlaps dynamically, depending on the global structure of the event's energy flow. (iv) ω_{cut} bounds the soft energy in the physically preferred totally inclusive fashion (cf. ref. [4]. (v) OJD is purely analytical, allowing its algorithmic implementations to differ beyond programmatic code optimizations and to be customized for specific applications. (vi) OJD is embedded in a systematic theory with new options for constructing improved data processing procedures that go beyond the conventional approach.

5. Despite the huge dimension of the domain in which to search the global minimum, $N_{\text{part}} \times N_{\text{jets}} = O(100–1000)$, OJD lends itself to efficient algorithmic implementations (the Optimal Jet Finder library [22]).

OJF was first developed in the programming language Component Pascal [24], featuring a unique combination of safety and efficiency. This was very useful for the experimentation needed to find a satisfactory algorithm. Only after that the final port to FORTRAN was performed. A subsequent testing [25] and a substantially independent verification [26] revealed no defects of significance, indicating a high reliability of the resulting code [23].

The OJF library can be used to obtain OJD implementations adapted for specific applications (see below).

6. A number of successive recombination algorithms were compared in ref. [9] in a series of tests none of which, however, was conclusive. The JADE algorithm proved to be the least satisfactory, the Geneva algorithm behaved somewhat erratically, and a group of algorithms (including k_{T} and Luclus) exhibited a balanced behavior in various tests, typically populating the spread between the JADE and Geneva algorithms. Note that the successive recombination scheme is recovered within OJD as a heuristic minimum-search trick with $n = 1$ in eq. 3 [17], which is the geometric mean of the JADE and Geneva criteria. Then OJD should roughly fall into the same group as the k_{T} and Luclus algorithms. A conclusive physically meaningful comparison can be performed in the context of the method of optimal observables. We explain the procedure using a simple example modeled after the measurements of the W-boson mass M at LEP2 [27]. The details will be published separately [28].

The process $e^+e^- \to W^+W^- \to$ hadrons at CM energy of 180 GeV was simulated using PYTHIA 6.2 [29]. Each event was resolved into 4 jets. These can be combined into two pairs (supposedly resulting from decays of the W's) in three different ways; we chose the combination with the smallest difference in invariant masses between the two pairs and calculate the average m of the two masses. This mapped events to the m axis. We used $9 \cdot 10^6$ events to generate the probability distribution $\pi_M(m)$ and to construct a numerical approximation to the optimal observable $f_{\mathrm{opt}}(m) = \partial_M \ln \pi_M(m)$. Using this as a generalized moment with a sample of N_{exp} experimental events would yield an estimate for M with the theoretically smallest error estimated as $\delta M_{\mathrm{exp}} \cong \left(N_{\mathrm{exp}} \left\langle f_{\mathrm{opt}}^2 \right\rangle\right)^{-1/2}$, where $\left\langle f_{\mathrm{opt}}^2 \right\rangle$ is sometimes identified with Fisher's information. δM_{exp} immediately reflects suitability of the jet algorithm used.

We thus compared OJD with the k_{T} and JADE definitions. We used the KTCLUS implementation of the k_{T} algorithm [30] and modified the recombination criterion to obtain the JADE algorithm. All events were forced to 4 jets, so the parameters y_{cut} and ω_{cut} played no role.

For OJD, we chose $R = 2$ and, for benchmarking purposes, first employed a primitive variant of OJF-based algorithm with a fixed n_{tries} for all events, where n_{tries} is the number of independent attempts to descend into a global minimum from a random initial configuration. The probability to miss the global minimum vanishes for larger n_{tries}; we chose $n_{\mathrm{tries}} = 10$. The obtained $f_{\mathrm{opt}}(m)$ for the three jet algorithms are shown in fig. 1.

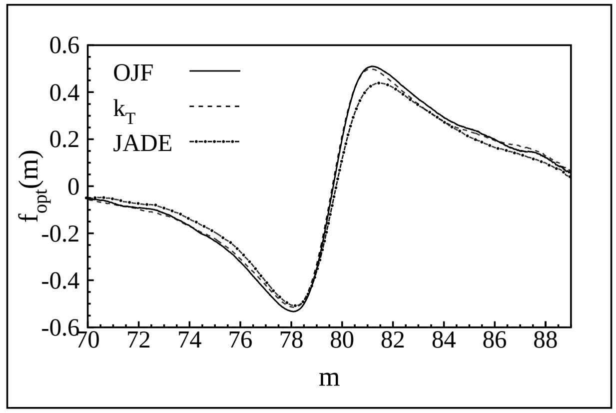

FIG. 1: Optimal observable $f_{\mathrm{opt}}(m)$ for OJF, k_{T} and JADE.

For $N_{\mathrm{exp}} = 1000$ (which roughly corresponds to the W-mass measurements at LEP2) we found the following:

ALGORITHM	$\delta M_{\mathrm{exp}} \pm 3$ MeV
OJD/OJF	106
k_{T}	105
JADE	118

The error of 3 MeV is mostly due to numerical differentiation in M.

Note that there are options to improve the measurement procedure that are specific to OJD, e.g. weighting events according to the values of Ω_R. We have not explored them, as it is sufficient for the purposes of this Letter to establish that OJD is at least no worse than the k_{T} algorithm for this measurement.

7. An important aspect is the speed of jet algorithms at large N_{part}. This is critical e.g. in the preclustering for reducing the number of clusters in each event as seen e.g. by the D0 detector at FNAL to about 200: otherwise it is not possible to analyze data with k_{T} as its processing time per event is $O\left(N_{\mathrm{part}}^3\right)$ [31]. A concern then is how the preclustering affects the final results as it has to be done using a method unrelated to the k_{T} algorithm, and a non-programmatic modification of the latter must be treated as a new jet definition (cf. examples in ref. [9]).

Speed of the algorithms as different as OJF and KTCLUS (coded in the same dialect of FORTRAN) may depend on the computing installation. With this in view, we report times per event in units of 10^{-2} sec as measured on our hardware with our sample of events.

N_{part} varied from 50 to 170 in our sample, with the mean value of 83. The processing time per event is described rather well by the following formulae:

$$\begin{array}{ll} 1.2 \times 10^{-6} \times N_{\mathrm{part}}^3 & \text{for KTCLUS,} \\ 1.0 \times 10^{-2} \times N_{\mathrm{part}} \times n_{\mathrm{tries}} & \text{for OJF.} \end{array} \tag{10}$$

This behavior was verified for N_{part} up to 1700 by splitting each particle into 10 collinear fragments (similarly to how a particle may lit up several detector cells). The required n_{tries} only depends on the number of local minima of Ω_R that reflects the event's global structure (number, width of jets, etc.) but not on N_{part}.

The simplest OJF-based implementation of OJD with a fixed n_{tries} for all events is faster than KTCLUS for $N_{\mathrm{part}} > 90\sqrt{n_{\mathrm{tries}}}$. Note that the values above 7 for n_{tries} seem to be rarely warranted, and for a substantial fraction of events very low values are in fact sufficient. We have not explored this option, focusing instead on a more significant optimization described below.

8. It is important to appreciate that whereas any modification of the k_T algorithm beyond an equivalent code transformation would have to be treated as an entirely new jet definition, OJD is formulated without reference to any specific implementation, so once a reliable minimization algorithm is found, it can be used to control the quality of other implementations designed for speed.

Useful modifications result from allowing a misidentification of the global minimum for a fraction of events, with the quality of the entire data processing procedure controlled via Fisher's information $\langle f_{\rm opt}^2 \rangle$. A simple such optimization can be implemented entirely using the routines from the OJF library; it relies on the well-known fact that the jet structure is often determined by the most energetic particles: Select the most energetic particles (a skeleton event), and precluster them by running the minimization routine. Then add the remaining particles with random values of z_{aj} and run the minimization again. With a threshold of 2 GeV to select the energetic particles, $n_{\rm tries} = 5$ at the preclustering phase and $n_{\rm tries} =1$ at the final stage, only a 1% change was observed for $\delta M_{\rm exp}$ (curiously, an improvement) whereas the speed much increased, with the dependence of the time per event on $N_{\rm part}$ now given roughly by

$$2.5 \times 10^{-2} \times N_{\rm part} \qquad (11)$$

with a hint at a slower growth at large $N_{\rm part}$. This is faster than KTCLUS starting from $N_{\rm part} \approx 140$, and the speed advantage increases sharply for higher $N_{\rm part}$: for $N_{\rm part} \approx 200$ this is twice as fast as KTCLUS, and an extrapolation to $N_{\rm part} \approx 1000$ yields the factor of 50.

The dramatically better behavior of OJF at large $N_{\rm part}$ makes it a candidate for work at the level of detector cells, perhaps even on-line (note that all $n_{\rm tries}$ minimization attempts can be done in parallel).

The OJF library implements the first minimization algorithm found to run acceptably fast. Better algorithms may be found once the OJD/OJF is explored further.

9. To summarize, a conclusive method to compare jet algorithms is based on evaluation of Fisher's information. In the considered model measurement, OJD is equivalent to the k_T definition in physical quality, and an implementation of OJD is increasingly faster than KTCLUS at large $N_{\rm part}$ starting from $N_{\rm part} \approx 140$. Moreover, OJD is defined in a theoretically preferred fashion and is supported by a systematic theory with new options for improvement of jets-based measurements. All this positions OJD as a candidate for a standard jet definition for the next generation of HEP experiments.

We thank A. Czarnecki and the two referees for useful criticisms. FT thanks A. Czarnecki for hospitality at the University of Alberta (Canada) where a part of this work was done. This work was supported in parts by the Natural Sciences and Engineering Research Council of Canada and the NATO grant PST.CLG.977751.

[1] F. V. Tkachov, Int. J. Mod. Phys. **A17**, 2783 (2002).
[2] See e.g. R. Barlow, Rep. Prog. Phys. **36**, 1067 (1993).
[3] F. Dydak, talk at The IX Int. Workshop on High Energy Physics (September 1994, Zvenigorod, Russia).
[4] G. Sterman and S. Weinberg, Phys. Rev. Lett. **39**, 1436 (1977).
[5] N. A. Sveshnikov and F. V. Tkachov, Phys. Lett. **B382**, 403 (1996).
[6] S. D. Ellis et al., in: Research Directions for the Decade, Snowmass 1990. Singapore: World Scientific, 1992.
[7] J. B. Babcock and R. E. Cutkosky, Nucl. Phys. **B176**, 113 (1980).
[8] S. Brandt et al., Phys. Lett. **12**, 57 (1964); E. Farhi, Phys. Rev. Lett. **39**, 1587 (1977).
[9] S. Moretti, L. Lonnblad and T. Sjostrand, JHEP **9808**, 1 (1998).
[10] T. Sjostrand, Comp. Phys. Comm. **28**, 229 (1983).
[11] JADE collaboration, Z. Phys. **C33**, 23 (1986).
[12] S. Bethke et al., Nucl. Phys. **B370**, 310 (1992).
[13] S. Catani et al., Phys. Lett. **269B**, 432 (1991).
[14] S. Youssef, Comp. Phys. Comm. **45**, 423 (1987).
[15] E. L. Berger et al., e-Print hep-ph/0201146.
[16] F. V. Tkachov, Phys. Rev. Lett. **73**, 2405 (1994); Erratum, **74**, 2618 (1995).
[17] F. V. Tkachov, Int. J. Mod. Phys. **A12**, 5411 (1997).
[18] F. V. Tkachov, Part. Nucl., Letters **2[111]**, 28 (2002).
[19] N. Amos et al., contribution to CHEP95, URL: http://www.hep.net/chep95/html/papers/p155/.
[20] P. C. Bhat, H. Prosper, and S. S. Snyder, Int. J. Mod. Phys. **A13**, 5113 (1998).
[21] C. F. Berger et al., e-Print hep-ph/0202207.
[22] D. Yu. Grigoriev and F. V. Tkachov, e-Print hep-ph/9912415; E. Jankowski, D. Yu. Grigoriev and F. V. Tkachov, to be submitted to Comp. Phys. Comm.
[23] The FORTRAN code ver. OJF_014 is publicly available from http://www.inr.ac.ru/~ftkachov/projects/jets/.
[24] http://www.oberon.ch.
[25] The first realistic test was run by P. Achard (L3, CERN) in 1999 with a sample of about 10^5 events.
[26] F. V. Tkachov, e-Print hep-ph/0111035.
[27] Technical report CERN-EP-2000-099.
[28] E. Jankowski and F. V. Tkachov, in preparation.
[29] T. Sjostrand et al., Comp. Phys. Comm. **135**, 238 (2001).
[30] http://hepwww.rl.ac.uk/theory/seymour/ktclus/
[31] Run II Jet Physics, e-Print hep-ex/0005012v2, Sec. 4.3.2.
[32] P. W. Bopp, Z. Phys. **C3**, 171 (1979).
[33] W. Bartel at al. (JADE Collab.), Phys. Lett. **B91**, 142 (1980).